



THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

# Detection of Mammalian Virulence Determinants in Highly Pathogenic Avian Influenza H5N1 Viruses: Multivariate Analysis of Published Data

### Citation for published version:

Lycett, SJ, Ward, MJ, Lewis, FI, Poon, AFY, Pond, SLK & Leigh Brown, A 2009, 'Detection of Mammalian Virulence Determinants in Highly Pathogenic Avian Influenza H5N1 Viruses: Multivariate Analysis of Published Data', *Journal of Virology*, vol. 83, no. 19, pp. 9901-9910. <https://doi.org/10.1128/JVI.00608-09>

### Digital Object Identifier (DOI):

[10.1128/JVI.00608-09](https://doi.org/10.1128/JVI.00608-09)

### Link:

[Link to publication record in Edinburgh Research Explorer](#)

### Document Version:

Early version, also known as pre-print

### Published In:

Journal of Virology

### Publisher Rights Statement:

Freely available via Pub Med.

### General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



## Detection of Mammalian Virulence Determinants in Highly Pathogenic Avian Influenza H5N1 Viruses: Multivariate Analysis of Published Data<sup>∇†</sup>

S. J. Lycett,<sup>1\*</sup> M. J. Ward,<sup>1</sup> F. I. Lewis,<sup>1‡</sup> A. F. Y. Poon,<sup>2</sup>  
S. L. Kosakovsky Pond,<sup>2</sup> and A. J. Leigh Brown<sup>1</sup>

*Institute of Evolutionary Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, United Kingdom,<sup>1</sup> and  
University of California at San Diego, 150 West Washington Street, Suite 100, San Diego, California 92103<sup>2</sup>*

Received 24 March 2009/Accepted 10 July 2009

**Highly pathogenic avian influenza (HPAI) virus H5N1 infects water and land fowl and can infect and cause mortality in mammals, including humans. However, HPAI H5N1 strains are not equally virulent in mammals, and some strains have been shown to cause only mild symptoms in experimental infections. Since most experimental studies of the basis of virulence in mammals have been small in scale, we undertook a meta-analysis of available experimental studies and used Bayesian graphical models (BGM) to increase the power of inference. We applied text-mining techniques to identify 27 individual studies that experimentally determined pathogenicity in HPAI H5N1 strains comprising 69 complete genome sequences. Amino acid sequence data in all 11 genes were coded as binary data for the presence or absence of mutations related to virulence in mammals or nonconsensus residues. Sites previously implicated as virulence determinants were examined for association with virulence in mammals in this data set, and the sites with the most significant association were selected for further BGM analysis. The analyses show that virulence in mammals is a complex genetic trait directly influenced by mutations in polymerase basic 1 (PB1) and PB2, nonstructural 1 (NS1), and hemagglutinin (HA) genes. Several intra- and intersegment correlations were also found, and we postulate that there may be two separate virulence mechanisms involving particular combinations of polymerase and NS1 mutations or of NS1 and HA mutations.**

H5N1 avian influenza initially garnered widespread attention following the first human cases of infection recorded in Hong Kong in 1997, after an outbreak in chickens (15, 72, 74), which itself followed an earlier infection in water fowl (21, 84). Since then the virus has spread across Asia to Russia, the Middle East, Europe, and Africa, causing deaths in wild aquatic birds (13, 58, 71, 81), domestic ducks and chickens (29, 39, 43), dogs and cats (2, 35, 36, 55, 76), and 254 of 405 confirmed human cases as of 5 February 2009 according to the World Health Organization ([http://www.who.int/csr/disease/avian\\_influenza/](http://www.who.int/csr/disease/avian_influenza/)).

Avian influenza virus strains are classified as highly pathogenic or of low pathogenicity on the basis of their virulence in chickens—highly pathogenic avian influenza (HPAI) virus can cause significant mortality (75 to 100%) in unvaccinated flocks (see reference 1), whereas low-pathogenicity strains cause only mild or no symptoms. HPAI H5N1 strains possess multiple basic amino acids in the hemagglutinin (HA) surface glycoprotein (27) cleavage site, the characteristic of all HPAI (62,

63). A variety of substrains of HPAI H5N1 have appeared during the last 10 years (82) and have probably become endemic in both wild birds and poultry (12, 43, 58). All strains of HPAI H5N1 are highly pathogenic in chickens, but their virulence varies in ducks (29) and mammals (see for example, references 11, 20, and 34).

The influenza A virus has an eight-segment RNA genome encoding 11 proteins, and the effects of point mutations or allelic combinations on the virulence of HPAI H5N1 isolates in mammals have been measured in mammals by reverse genetics and reassortment studies. Virulence in mammals is thought to be polygenic (10, 20), and several experimental studies have shown that mutations in the genes encoding internal proteins are important virulence determinants. For example, E627K in PB2 was shown to be important for replication in mammalian cells (73) and in the difference in virulence of two H5N1 Hong Kong 1997 outbreak isolates (24), whereas D92E in NS1 was found to increase resistance to tumor necrosis factor alpha and gamma interferon host responses for another H5N1 Hong Kong 1997 strain in vitro and in vivo in one study in swine (67).

Given the diversity of H5N1 viruses, their high rates of mutation and reassortment (14, 21, 22, 68, 77), and the polygenic nature of their virulence in mammals, general conclusions about virulence factors cannot be drawn with confidence from individual studies of small numbers of closely related sequences. However, systematically amalgamating the results from several studies may allow a more comprehensive picture to emerge. Such integration of results from many experiments and assays to provide a more complete understanding of gene-protein interactions has been carried out in *Saccharomyces*

\* Corresponding author. Mailing address: Ashworth Laboratories, Rm. 65, Institute of Evolutionary Biology, University of Edinburgh, Kings Buildings, West Mains Road, Edinburgh EH9 3JT, United Kingdom. Phone: 44 131 650 8683. Fax: 44 131 650 5456. E-mail: samantha.lycett@ed.ac.uk.

† Supplemental material for this article may be found at <http://jvi.asm.org/>.

‡ Present address: Epidemiology Research Unit, Drummondhill, Stratherrick Road, Scottish Agricultural College, Inverness IV2 4JZ, United Kingdom.

<sup>∇</sup> Published ahead of print on 22 July 2009.

*cerevisiae* (32, 40, 65). Using multiple data sources reduces the number of false-positive results and can reveal new patterns of association (32, 40), provided data quality and content issues are overcome (see for example, reference 56). Furthermore, while assembly of a few large data sets into a form suitable for meta-analysis can be done manually, experimental results from many small individual studies can also be captured by text mining published abstracts from online databases such as PubMed at the National Center for Biotechnology Information (NCBI) (51, 69).

The aim of the present study was to find a statistical model to describe the amino acid sites important for virulence of HPAI H5N1 in mammals and to indicate possible interactions between the sites. We have text mined PubMed abstracts for H5N1 isolates with experimentally determined virulence in mammals and downloaded the corresponding full genome sequences from the NCBI Influenza Virus Resource (4). We then fitted a Bayesian graphical model (BGM) to capture the dependencies among mutations and between individual mutations and the mammalian virulence phenotype. BGMs represent the probabilistic dependencies between random variables as a network without self-references, i.e., a directed acyclic graph (37, 59). They provide a convenient description of multivariate data because each link encodes a direct probabilistic dependency between sampled variables (e.g.,  $A - C$ ), rather than just a correlation (e.g.,  $A - C$ , because  $A - B$  and  $B - C$ ). BGMs have been used in several biological contexts: to infer gene expression networks from microarray data (18), to find the genes involved in a complex trait using single nucleotide polymorphism data (66), and to find coevolving sites in a rapidly variable region of a human immunodeficiency virus (HIV) gene (61). In the present study the nodes (random variables) in the network represent variable amino acid sites or the phenotype: links between nodes represent direct dependencies, i.e., links between amino acid sites represent putative functional interactions, and links between amino acid sites and the phenotype of “virulence in mammals” represent direct association of specific mutations and virulence.

## MATERIALS AND METHODS

**Text mining.** A list of H5N1 HPAI isolates with experimentally determined virulence in mammals was obtained by using a text-mining approach, followed by manual curation. Initially, XML (Extensible Markup Language) records of all abstracts containing “H5N1” and “virulence” keywords were obtained from PubMed using Eutils queries ([http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils\\_help.html](http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html)) generated with R scripts (<http://www.r-project.org>). Each of the 542 abstracts obtained on 5 March 08 were searched for isolate names (such as A/Vietnam/1203/2004) using a “/(words)/(words)” regular expression (i.e., at least three “/”s). This yielded 185 isolate names in 97 abstracts. Next, a list was compiled (using R scripts and manual editing) of unique isolate names, taking into account synonyms (e.g., dk = duck) and recording the PubMed identifiers (PMIDs) of papers that had mentioned the strain. Of the 94 unique isolate names found, 64 were for H5N1 isolates with measured virulence in mammals and full genome sequences deposited in the NCBI Flu database (4) (<http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html>).

The abstracts corresponding to the 64 H5N1 isolates were examined, and detailed results from mice or ferret studies were manually extracted from the corresponding studies. In some studies additional strains not mentioned in the abstract had also been studied, and these strains were also added to the original list, bringing the total to 69 strains from 27 individual studies. Table SA1 in the supplemental material contains the references to the studies and the numbers of strains from each study in the final data set. Each strain was manually classified as either virulent or nonvirulent in mammals based upon the amalgamated

experimental evidence. Table SA2 in the supplemental material contains the assigned classification and experimental evidence per strain.

**Data preprocessing.** Full genome nucleotide sequences of the 69 isolates with experimental evidence for (or against) virulence in mammals were downloaded from the NCBI Flu Database. The individual segments were initially aligned by using CLUSTAL W in BioEdit, and then a final manual codon alignment was performed in MEGA 4.0. The PB1-F2, M2, and NS2 coding sequences were manually extracted from the aligned nucleotide files (not all strains were annotated for these alternative coding sequences), and all 11 nucleotide coding sequences were translated to protein sequences in MEGA 4.0.

To analyze the association of particular residues at amino acid sites and virulence using BGMs, the protein sequences were converted into a binary matrix where each row of the matrix represented an isolate genome (11 proteins concatenated), and columns represented amino acid sites. An additional column for the virulence phenotype was added. Zeros represented consensus amino acid residues, ones represented any mutant (or the virulent phenotype). In particular, the following procedure was used to determine the binary coding at each variable site (via a custom R script). (i) If there are only two types of amino acids present at a site ( $A$  or  $B$ ), a  $2 \times 2$  table was calculated, counting the number of  $A$  residues in the virulent sequences,  $V(A)$ ; the number of  $A$ 's in nonvirulent sequences,  $N(A)$ ; the number of  $B$ 's in virulent sequences,  $V(B)$ ; and the number of  $B$ 's in nonvirulent sequences,  $N(B)$ . (ii) To determine which to code as 1,  $A$  or  $B$ , the determinant of the  $2 \times 2$  table was calculated as  $V(A)N(B) - V(B)N(A)$ . If the determinant is positive, then amino acids of type  $A$  are coded as 1's and the  $B$ 's are coded as 0's (and vice versa for negative determinants). Alternatively, if there is evidence in the literature to indicate that a particular residue is related to virulence (or nonvirulence), then this residue is coded as 1 (or 0). (iii) If there are more than two types of residue at a site, then step 2 is repeated for all combinations of amino acids partitioned into two groups, and the combination with the smallest  $P$  value from the Fisher exact test on the  $2 \times 2$  table is chosen.

The coding scheme used, corresponding  $2 \times 2$  table entries and  $P$  value from the Fisher exact test for sites of interest are reported in Table 1 (see Table SA3 in the supplemental material for information on all of the sites examined).

**Multivariate analysis of variable amino acids.** (i) **BGMs.** BGMs were inferred by using HyPhy (60, 61) from sequence and phenotype data coded to a binary matrix. HyPhy implements the methods of Cooper and Herskovits (16) and Friedman and Koller (19) for BGM inference as follows. The probability for a particular graph structure given the data can be calculated by using the K2 scoring metric (16). The K2 metric score for one node is a function of the number of discrete states of the node and its parents (e.g., how many  $B = 1$  when  $A = 1$ , etc.), and the total score for the graph is the product of the individual node scores. When the number of nodes is nontrivial (typically greater than 5), evaluating all possible graph structures becomes prohibitive so the Monte Carlo Markov chain (MCMC) method is used over families of ordered nodes to find a set of probable graph structures to describe the data (19). The output of the BGM inference is a consensus graph, where each link is assigned a probability of existing according to how often it appears in the sampled MCMC results (i.e., the marginal posterior probability – the expectation of posterior probabilities weighted by the likelihoods of the node orderings in which the given link appears). Links with probability greater than or equal to 0.5 in the consensus graph are used to create the final network.

For analysis of the 69 sequence data set (with 10 nodes), the HyPhy BGM MCMC was run with  $10^6$  steps (after a burn-in of  $10^5$  steps) and sampled every 1,000 steps. These parameters were chosen so that the sampled likelihood values were converged to a stationary distribution, and the individual samples were independent (the autocorrelation of the sampled likelihood values was calculated in every run, and the values of the first lags were not significantly different from zero). The maximum number of parents allowed per node was increased until there was no significant change in network structure (i.e., no new links where  $P \geq 0.5$  occurred), and a maximum of five parents per node was found to be sufficient for the present data set.

(ii) **Model validation.** The significance of the links in the BGM inferred from the  $69 \times 10$  binary data matrix was tested in two ways.

(a) The inferred BGM was compared to a set of null model BGMs. Null models were generated by inferring BGMs from nonparametrically bootstrapped data in which one site (or phenotype) was randomly permuted (i.e., a fixed column in a matrix of observation was resampled without replacement) at a time (leaving the other sites unchanged). Ten randomizations were performed per site (i.e., one-hundred randomizations for the ten sites). Randomizing each site (or the phenotype) independently of the others and inferring the BGM gives the background probability of the links between that site and the others due to random noise alone, conditional upon the joint probabilistic structure of the remaining data. The unpermuted BGM link probabilities from the site of interest

TABLE 1. Amino acid site coding and *P* value for association with virulence using the Fisher exact test<sup>a</sup>

Site	Residue(s)		<i>P</i>	Selected	Reference(s)	Additional detail(s)
	Virulent	Nonvirulent				
PB2-318	K	R	0.0027	*	10	
PB2-355	K	R, Q	0.0003	(PB1-317)	10, 34, 41	
PB2-627	K	E	0.0034	*	10, 11, 24, 41	
PB2-627 or 701	K, N	E, D	0.0003	(*)	17, 38, 70	
PB1-317	I	M, V	3.8E-05	*	10, 34, 41	
PA-127	I	V	0.0007	(PB2-318)	75	
PA-336	M	L	0.0006	(PB2-318)	75	
HA-102 (86E)	V	A, I, P, S, T	0.0092	*	83	
HA-140 (124B)	S	N, D	0.0431		83	
HA-154 (138A)	L, N	Q, H, I	0.0391		83	
HA-172 (156 glyco)	T, S	A	0.0010	*	8, 10, 83	
HA-228 (212D)	E, R	K	0.0194	*	83	
HA-279 (263E)	T	A	1.8E-05	*	83	
NA-49:72	Any deletion	No deletion	0.0003	*	5, 52	Stalk deletion
NS1-42	S	A, P	0.0244		33	P attenuates pathogenicity
NS1-92	E	D	0.0010	*	41, 67	
NS1-92/97	E	D	0.0504		48	D92E (no deletion) or D97E (with deletion)
NS1-127	N	T, D, R, V, A	0.0387		53	123-127 PKR region
NS1-189	N	D, G	0.0006	(PB2-318)	75	
NS1-195	T, Y	S	0.0294	(HA-102)	7	
NS1-228	P	S	0.0027	(NS1-92)	31, 57	PDZ binding domain
NS2-31	I	M	0.0006	(PB2-318)	75	
NS2-56	Y	H, L	0.0006	(PB2-318)	75	

<sup>a</sup> For each sequence, sites containing the “virulent” residues (column 2) were coded as ones or zeros (column 3). The numbers of sequences classified as virulent (nonvirulent) with virulent (nonvirulent) residues were calculated, and the *P* value from the Fisher exact test (uncorrected) is displayed in column 4. The nine sites with the lowest *P* values selected for further analysis are indicated by an asterisk in column 5; sites with the same (or similar) pattern of mutations to selected sites are also indicated in column 5. All site numbering is from the first methionine in the coding sequences. The numbers in parentheses for HA correspond to the mature H5 site numbers, as in reference 83, and the letters correspond to the canonical antigenic site, or a glycosylation site (glyco). Abbreviations: PB, polymerase basic; NP, nucleoprotein; NA, neuraminidase; M, matrix protein; NS, nonstructural protein.

to or from the other sites were normalized by the mean and standard deviation of the 10 permuted results to give a standard score per link.

(b) Cross-validation of structure was applied. To test the robustness of the inferred BGM and mitigate any possible effects from isolate misclassification or sequencing errors, 10 BGMs were calculated and averaged over different subsets of the data. Each subset of data consisted of 65 sequences (i.e., training data), and the 4 remaining sequences were kept as test sets. Since a different group of strains (particularly the Z constellation) began to replace the early strains from 2002 onward (11, 22, 43), different patterns of mutations may occur after 2002; hence, four sequences were excluded in each run, one from each of the pre-2002 (early) virulent, 2002+ (late) virulent, pre-2002 nonvirulent, and 2002+ nonvirulent sequences.

(iii) **Conditional probability tables.** Conditional probability tables describing the probability that a strain is virulent (nonvirulent) given the presence or absence of mutations at individual sites (identified in the BGM as directly influencing virulence), or a combination of those sites, were calculated separately for each of the 10 cross-validation training data sets using the observed frequency of mutations. Certain combinations of mutations in the influencing sites were not observed because of the correlation between the sites (e.g., there were no sequences in the 69 sequence set with PB1-317 = 1 and NS1-92 = 0 or vice versa). For (hypothetical) strains with these unobserved combinations, we estimated the probability of virulence from the probability of virulence of the individual sites, assuming that the sites were independent as in equation 1 below.

$$P(\text{Vir} = 1|a_1a_2a_3) = 1 - \prod_{i=1}^3 [1 - p_i(a_i)] \quad (1)$$

where  $P(\text{Vir} = 1|a_1a_2a_3)$  is the probability of virulence for a sequence with the particular combination  $a_1a_2a_3$  of amino acids, and  $p_i(a_i)$  is the probability of virulence when node  $i$  has amino acid  $i$ .

The performance of the model was examined by predicting the virulence of the sequences in each of the 10 training and test data set pairs using the conditional probability table derived from each training set, respectively. The numbers of true positives (*TP*; virulent sequences predicted to be virulent by the model), false negatives (*FN*; virulent sequences predicted as nonvirulent), false positives

(*FP*; nonvirulent sequences predicted as virulent), and true negatives (*TN*; non-virulent sequences predicted as nonvirulent) were calculated along with the misclassification ( $FP + FN$ ), sensitivity [ $TP/(TP + FN)$ ], and specificity [ $TN/(FP + TN)$ ] values.

## RESULTS

**Isolates with experimentally determined virulence in mammals.** A total of 69 H5N1 full-genome isolates from 27 studies of virulence in mammals identified using a text-mining approach, followed by manual curation, were used in this analysis (see Materials and Methods and Table SA1 in the supplemental material). The most frequently measured strain—A/Vietnam/1203/04—was the subject of seven virulence studies. Five studies (11, 20, 34, 47, 50) contained measurements for at least five sequences. Strains were classified as virulent (or nonvirulent) if infection studies showed a high mortality with a low dose; sequences with 50% lethal dose of  $\leq 10^3$  50% egg infectious doses were classified as virulent in mammals. Even though the virulence results were obtained in different animals and using different protocols, the results were generally concordant: in 75% of cases (52 sequences), strains could be classified clearly either as highly pathogenic in mice or ferrets or only caused mild or no symptoms. Furthermore, three studies measured virulence in both mice and ferrets (28, 50, 64) in a total of five strains. In these cases, the pathogenicity in ferrets correlated with that in mice (highly pathogenic strains in ferrets were highly pathogenic in mice, low-pathogenicity strains in ferrets showed low pathogenicity in mice). Although differences in the virulence of two H5N1 strains in mice and ferrets



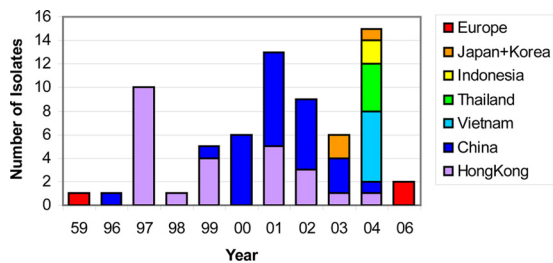


FIG. 1. Distribution of full-genome isolates in text-mined data across isolation year and region. The total number of isolates in the data set per year is indicated by the height of the bars, and the distribution by region is represented by the bar colors. The data covers three major outbreaks: Hong Kong in 1997, Hong Kong/China in 2001, Vietnam (plus others) in 2004 over a period of at least 10 years.

at high dose have been reported (86), we did not find significant differences in the overall virulence of the strains in the present data set when we used our binary classification scheme. Details of the strain classification and associated experimental evidence can be found in Table SA2 in the supplemental material.

The final full data set consisted of 69 full genome sequences: 35 from water fowl, 17 from land fowl, and 17 from humans. Totals of 20% (7) of the water fowl isolates, 24% (4) of the land fowl isolates, and 65% (11) of the human sequences were classified as virulent in mammals, making a total of 22 virulent sequences. The distribution of the strains across isolation year and place is shown in Fig. 1: this data set covers several outbreaks, including the 1997 Hong Kong, 2000 to 2002 China, and 2004 Viet Nam epidemics.

**Site association with virulence.** Across the whole genome, 1,235 of 4,582 (~27%) amino acid sites were variable ( $\geq 1$  alternative residue in the 69 sequences) with 969 of 1,235 (~78%) of these variable sites containing only 2 amino acid residues. Furthermore, in the remaining 266 (22%) of variable sites that coded for three or more alternative amino acids, the average frequency of the next most abundant residue was only 4%. To facilitate analysis of the relationship between mutations and virulence in mammals, the sequence data were coded to a binary matrix, which therefore retained almost all of the information contained in the sequence alignment.

All 1,235 polymorphic amino acid sites in the protein sequences of the 69 strains were examined, and the distribution of the alternative residues among strains classified as virulent or nonvirulent was recorded. A total of 227 amino acid sites (19% of the variable sites) across the entire genome were found to have an association with virulence at  $P \leq 0.05$  (Fisher exact test, uncorrected). However, there was a high degree of association among sites, with mutations at 129 sites having an identical distribution across the strains to at least one other site. Of the 98 sites with a distribution that was unique in the data set, only 53 differed by more than one mutation from any other site. Of the 98 sites with a unique distribution, 12 have been reported in the literature to have functional significance for virulence (see Table SA3 in the supplemental material).

We next examined all amino acid sites or genomic features (e.g., neuraminidase [NA] stalk deletion; PDZ binding domain in NS1) previously reported in the literature as having functional significance for virulence in H5N1 and found evidence

for 70 sites/features. Their distribution among virulent and nonvirulent strains was tested as described above: the 23 that were apparently associated with virulence in this data set are shown in Table 1 (Fisher exact test  $P \leq 0.05$ , uncorrected; see Table SA3 in the supplemental material for details of all 70 sites). All of the strains in this data set, including nonvirulent ones, contained the HA-226(Q) and HA-228(G) (H3 numbering) residues for the avian sialic acid receptor binding site (23, 30) and the NS1-149(A) virulence determinant in chickens (45); on the other hand, only one strain contained the pathogenicity-associated proline at amino acid site 42 in NS1 (33).

Of the 23 sites, three in PB2 (sites 318, 355, and 627), one in PB1 (site 317) and one in NS1 (site 92) which were apparently associated with virulence ( $P < 0.01$ ; Table 1) had previously been claimed to have a role in virulence determination (see the Discussion). In addition, the presence of PB2-701N has been suggested as an alternative virulence or adaptation marker in mammals in place of PB2-627K (17, 38, 70), and combining the PB2-627 and PB2-701 sites resulted in a more significant  $P$  value ( $P = 0.0003$ ) for association with virulence than PB2-627 ( $P = 0.003$ ) or PB2-701 ( $P = 0.08$ ) alone. A further five sites in polymerase acidic (PA), and nonstructural genes (NS1 and NS2) were also nominally significant; however, these sites had very similar distributions across strains to PB2-318. Two HA antigenic sites—HA-102(mature H5 site 86, part of canonical antigenic site E) and HA-279(263E) and the glycosylation site HA-172(156)—had a possible association with virulence ( $P < 0.01$ ), and deletions in the NA stalk region appeared to be quite strongly associated: 21 of the 22 virulent sequences showed stalk deletion compared to about half of the nonvirulent sequences ( $P < 0.001$ ).

**Site selection for BGM.** The uncorrected associations are not themselves evidence of causal dependencies but assist with the selection of sites to be analyzed in the multivariate model. Sites with the same (or only one mutation different) pattern across the sequences were combined since the individual contribution of these sites to virulence would not be distinguishable from this data set. The sites with common patterns are shown in Table 2. Of the 23 sites and/or features with Fisher exact test  $P$  values of  $\leq 0.05$ , only 13 had distinct patterns. In order to reduce the possibility of inferring false links, the total number of variables included in the BGM analysis was restricted as indicated by simulation studies based on sample sizes of 60 to 70 strains (see the discussion of simulation analysis in the supplemental material). Consequently, the nine

TABLE 2. Amino acid sites with similar patterns of mutations<sup>a</sup>

Pattern	Sites (virulent nonvirulent)
1 .....	PB2-318(K R), PA-127(I V), PA-336(M L), NS1-189(N DG), NS2-31(I M), NS2-56(YL H)
2 .....	PB2-355(K RQ), PB1-317(I MV)
3 .....	PB2-675(L I), PB2-683(T A), PB1-198(K R), NA-39(QH K), NA-223(I T)
4 .....	HA-102(V AIPST), NS1-195(TY S)
5 .....	NS1-92(E D), NS1-228(P S)
6 .....	M2-64(SAF P), M2-69(P L)

<sup>a</sup> The distributions of mutations across strains for sites identified in the literature (see Table 1 and Table SA3 in the supplemental material) were examined. These groups of sites have more than one mutation and identical or nearly identical distributions ( $\leq 1$  mutation difference).

TABLE 3. Sites selected for BGM<sup>a</sup>

Site (virulent nonvirulent)	Comment(s)
Vir	"Virulent-in-mammals" phenotype
PB2-318(K R) ++	Combined pattern 1, see Table 2
PB1-317(I MV)/PB2-355(K RQ)	Combined pattern 2, see Table 2
PB2-627(K E)	
HA-102(V AIPST)/NS1-195(TY S)	Combined pattern 4, see Table 2
HA-172(TS A)	
HA-228(ER K)	
HA-279(T A)	
NA-del	NA stalk deletion ("1") or no deletion ("0")
NS1-92(E D)/NS1-228(P S)	Combined pattern 5, see Table 2

<sup>a</sup> The sites included in the BGM analysis were the nine sites with the most significant *P* values for association with virulence in mammals, and the "virulent-in-mammals" phenotype. Sites with the same distribution of mutations across the strains are analyzed together as indicated.

uniquely patterned sites with the most significant *P* values for association with virulence in mammals (ranging from  $P < 10^{-4}$  to  $P = 0.019$ ), together with the virulence phenotype, were included in the multivariate analysis (Table 3). The selected sites were: PB2-318, PB2-627, PB1-317 (≡ PB2-355), HA-102 (≡ NS1-195), HA-172, HA-228, HA-279, NA stalk deletion, and NS1-92 (≡ NS1-228) (the final binary data used can be found in Table SA4 in the supplemental material). Since there is evidence in the literature to suggest that PB2-701N may compensate for the lack of PB2-627K in the adaptation of the virus to mammals (17, 38, 70), we also investigated the effect of using the combined PB2-627 and PB2-701 site data in place of PB2-627 alone.

**BGM results.** BGMs were inferred from the binary data of the selected sites using: (i) all 69 sequences; (ii) 100 cases of permuted binary data (null models) and; (iii) 10 cases of leaving 4 sequences out each time (10× cross validation, see the discussion of model validation in Materials and Methods). Figure 2 shows the inferred network structure for significant links, together with the average link probability from the 10× cross validation and standard score from the comparison to the null model networks. The model confirms that virulence in mammals is a complex genetic trait affected directly by mutations in the polymerase, NS1, and HA genes.

Very strong correlations among certain sites leave some ambiguity over the precise nature of the molecular changes since these sites could not be analyzed separately (Table 2). Specifically, three distinct mutation patterns, representing two sites each—PB1-317/PB2-355, NS1-92/NS1-228, and HA-102/NS1-195—were directly associated with virulence in mammals in the model. We did not detect a significant direct association of PB2-627 alone with virulence in mammals in the model; however, a strong direct association, with a probability of  $\geq 0.8$  (standard score = 3 to 4) was detected for the combined sites PB2-627 and PB2-701 (data not shown).

Sites within the polymerase genes were strongly associated with each other in the model: the pattern 1 sites (including PB2-318, PA-127, and PA-336) were linked to PB1-317/PB2-355 pattern 2 sites with a probability of  $\geq 0.9$  (standard score = 9 to 10). Strong intrasegment associations were also found in both HA and NS, particularly between antigenic sites HA-102

and HA-279, HA-102 and the variable glycosylation site HA-172, and NS1-92 and NS1-228 (pattern 5) and between NS1-189, NS2-31, and NS2-56 (pattern 1). The most significant intersegment association was between PB1-317/PB2-355 (pattern 2) and NS1-92/228 (pattern 5). NS1 sites 92, 195, and 228 were also strongly associated with HA-279 (antigenic site), and NS1-189 had the same pattern as the PB2-318, PA, and NS2 sites (pattern 1). Additionally, the previously reported interaction between the HA-172(156) glycosylation site and the NA stalk deletion (3) was detected in the BGM with a probability of  $\geq 0.8$  (standard score = 9 to 10).

The model was tested by using cross-validation (see Materials and Methods). Table 4 and Table 5 show the conditional probabilities for virulence in mammals averaged over each training data case from the cross-validation data sets, calculated per single influencing node (Table 4) and per combination of influencing sites (Table 5). Where a particular combination of amino acids has not occurred in this data set (due to the small data set size and interactions between the influencing sites), the probability of virulence in Table 5 was estimated by combining the individual probabilities in Table 4, assuming independence among sites (see the discussion of conditional probability tables in Materials and Methods).

Using the conditional probability tables from the 10 training data examples, virulence predictions were made for each test data example (these were composed of the four excluded sequences from the training data: one of each early-

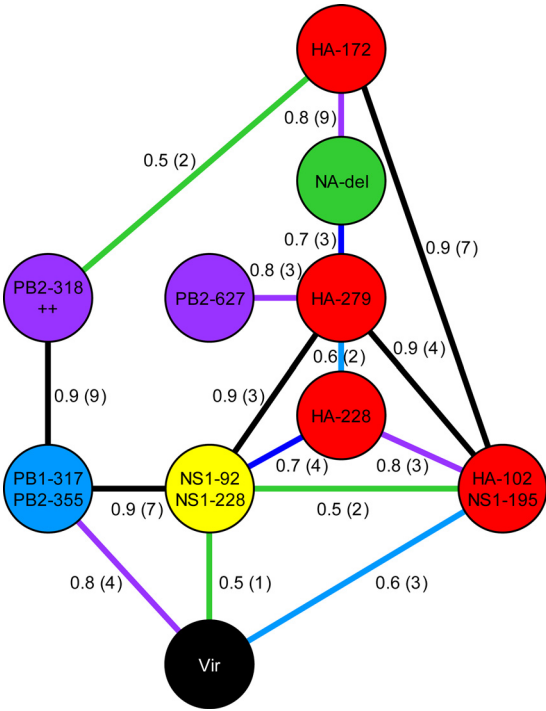


FIG. 2. Inferred BGM for virulence in mammals. The indicated link probabilities from 10× cross validation and z-score from comparison to null models (in parentheses) are rounded down, e.g., "0.8 (3)" means a link probability of 0.8 to 0.9 and a z-score of 3 to 4. Note that some nodes represent more than one site, as listed in Table 2. The sites directly associated with virulence in mammals are PB1-317/PB2-355, NS1-92/NS1-228, and HA-102/NS1-19.

TABLE 4. Conditional probabilities for “virulent-in-mammals” sequences given individual sites<sup>a</sup>

Node	Sites	Vir-AA	Non-Vir AA	$P(\text{virulence} \text{Vir-AA})$	$P(\text{virulence} \text{non-Vir AA})$
1	PB1-317/PB2-355	I/K	MV/RQ	1.00	0.22
2	NS1-92/NS1-228	E/P	D/S	0.73	0.22
3	HA-102/NS1-195	V/TY	AIPST/S	0.69	0.24

<sup>a</sup> Given that a sequence has a “virulent” type amino acid (Vir-AA) at a site (as indicated in column 3), the probability of that sequence having a “virulent” classification is indicated in column 5. If the sequence has a “nonvirulent” amino acid (Non-Vir AA) at a site (column 4), the probability of that sequence still having a “virulent” classification is indicated in column 6. For example, if the sequence contains PB1-317I (and/or PB2-355K), the probability of that sequence being virulent is 1; if the sequence does not have these mutations, then the probability of virulence is 0.22. All probabilities were calculated by averaging over the 10 training data examples and assuming independence between sites (rows).

virulent, late-virulent, early-nonvirulent, and late-nonvirulent sequences, where early refers to a pre-2002 occurrence; see the discussion of model validation in Materials and Methods for details). The average prediction performance over the cross validation training and test data sets was calculated (Table 6); for the test data, the sensitivity was 70%, and the specificity was 95%.

## DISCUSSION

This study combines results from previous experimental studies based on strains isolated during 3 major HPAI H5N1 outbreaks in order to examine the genetic basis of virulence in mammals. Extensive literature mining, followed by rigorous quality control, identified 69 complete genome sequences with experimentally determined virulence in mammals for use in this analysis, a threefold increase from the number of sequences analyzed in the largest experimental study to date. We have fitted for the first time a rigorous statistical model relating mutations in multiple influenza viral segments to virulence in mammals. Many authors have previously described virulence

TABLE 5. Conditional probabilities for “virulent in mammals” sequences for combinations of sites<sup>a</sup>

Node <sup>b</sup>			$P(\text{virulence} \text{state})^c$	Virulence prediction ( $P \geq 0.5$ )
1	2	3		
0	0	0	0.09	0
0	0	1	0.70	1
0	1	0	0.25	0
0	1	1	0.94*	1
1	0	0	1*	1
1	0	1	1*	1
1	1	0	1	1
1	1	1	1*	1

<sup>a</sup> A sequence with the combination of mutations [“virulent” (1)/“nonvirulent” (0) amino acids] as indicated in columns 1 to 3 has a probability of virulence in mammals as indicated in column 4. If the probability of virulence in mammals is  $>0.5$ , then the sequence is classified as “virulent”.

<sup>b</sup> Node 1, PB1-317/PB2-355; node 2, NS1-92/NS1-228; node 3, HA-102/NS1-195.

<sup>c</sup> The values marked with an asterisk indicate that this particular combination of mutations was not observed in the training data sets (or 69 sequence data set), so the probability of virulence in mammals was estimated from the individual values in Table 4 (see Materials and Methods). All probabilities were calculated by averaging over the 10 training data examples.

TABLE 6. Average prediction performance over cross-validation training and test data sets<sup>a</sup>

Data set	Avg performance (%)	
	Training	Test
TP	13.6 (20.9)	1.4 (35)
FN	6.4 (9.8)	0.6 (15)
FP	2.9 (4.5)	0.1 (2.5)
TN	42.1 (64.7)	1.9 (47.5)
MC	9.3 (14.3)	0.7 (17.5)

<sup>a</sup> Each sequence in the training and test data sets was classified as virulent or nonvirulent according to the combination of amino acids at PB1-317/PB2-355, NS1-92/NS1-228, and HA-102/NS1-195 as in Table 5. The average numbers (or percentages) of true positives (TP; virulent sequences classified as virulent), false negatives (FN), false positives (FP), and true negatives (TN) from the 10 training and test data sets, and the misclassification (MC) performance metrics are indicated. The sensitivity and specificity values were 68% and 94%, respectively, for the training data sets and 70% and 95%, respectively, for the test data sets.

in terms of a complex trait (10, 11, 20): the model presented here confirms that virulence in mammals is affected by mutations in at least four segments: PB1, PB2, HA, and NS1.

**Univariate analyses.** Three sites in PB2 (sites 318, 355, and 627), PB1-317 and NS1-92 that were apparently associated with virulence in the univariate studies had been proposed in earlier experimental studies (10, 11, 24, 34, 41, 50, 67). Recent experimental evidence indicates that a PB2-701N mutation may arise in mammals instead of PB2-627K (70); we found no isolates containing both mutations in the 69-sequence data set, and combining these two sites increases the association with virulence. A further five sites in PA, NS1, and NS2 identified in a study of the 1997 Hong Kong H5N1 human outbreak (75) were also nominally significant; however, these sites had very similar mutation patterns to PB2-318 and the “virulent” residues only appeared in the 1997 sequences. We did not detect any significant association of NS1 deletions alone with virulence in mammals in this data set (48). However, since NS1-97E is a conserved residue (in the aligned sequences), the 5-amino-acid NS1-deletion also leads to an E at position 92 in the protein (48), and NS1-92E (no deletion) or NS1-97E (with deletion) with virulence in mammals was marginally significant ( $P \sim 0.05$ ). The association of NS1-92E alone (aligned sequences) with virulence in mammals was more significant in this data set ( $P < 0.001$ ). In addition, no significant associations between virulence in mammals and the Src homology 3 domain binding motifs or PDZ binding domains containing the amino acid motif ESEV or EPEV in NS1 were found (26, 31, 57). However, proline at NS1-228 (part of the PDZ binding domain) was associated with both virulence in mammals ( $P < 0.003$ ) and strongly with E at NS1-92 ( $P \ll 10^{-6}$ ).

**Multivariate BGM models.** The nine sites with the strongest individual associations with virulence were included in the BGM analysis, and three of them (PB1-317/PB2-355, NS1-92/NS1-228 and HA-102/NS1-195) were found to have a direct association with virulence.

The mutation patterns at PB1-317 and PB2-355 were very similar to each other in the 69-sequence set and directly associated with virulence in mammals. These mutations were first identified by Katz et al. (34) as correlated with virulence in mice in the 1997 Hong Kong outbreak. In this 69-sequence data set, the PB1 (M,V)317I and PB2 (R,Q)355K “virulent”



mutations only appear in sequences from the 1997 Hong Kong outbreak. In an alignment of 579 full-genome sequences (representing all H5N1 unique sequences deposited in the NCBI Influenza Virus Resource as of September 2008), 44 sequences contained one and 12 contained both of the mutations. Apart from the 1997 Hong Kong outbreak, PB1-317I and/or PB2-355K mutations were also present in sequences from birds in China and Viet Nam in 2005 and 2006; a civet (Viet Nam 2005); and four human sequences (Indonesia 2006, China 2006 and 2007).

Two mutations in NS1, D92E and S228P, which were strongly associated in the data set were also directly associated with virulence in mammals in the BGM. The NS1 D92E change is thought to influence pathogenicity by making the virus more resistant to host interferon and tumor necrosis factor alpha responses (67), thus inducing a more severe cytokine response in the host (46). NS1-228 is one of the C-terminal residues and part of a region that could bind proteins containing PDZ domains. NS1 proteins with avian ESEV or EPEV PDZ domain ligand motifs were postulated to disrupt human cell pathways by binding to proteins with PDZ domains (57), and NS1 proteins with avian PDZ domain ligands have been shown to cause increased virulence in mice (31). A total of 51 of the 69 sequences examined here contained ESEV or EPEV at positions 227 to 230: no significant association of these motifs with mammalian virulence was found. However, the point mutation S228P was significantly associated with virulence in mammals. The NS1-92 and NS1-228 mutations remain correlated in the 579 H5N1 full genome sequence set. There are 17 sequences with both mutations, the majority were from the 1997 Hong Kong outbreak, but 3 were from isolated from water fowl in Viet Nam 2005. Five sequences (one from swine in China 2003 and four from birds in Viet Nam and Russia 2005) contained the S228P mutation only, but no sequences contained only the D92E mutation. The BGM inferred from the 69-sequence data set showed a strong link between PB1-317/PB2-355 and NS1-92/NS1-228, and this association still holds in the 579 full-genome data set ( $P \ll 10^{-6}$  for the Fisher Exact test with the Bonferroni correction for multiple testing between any of the 2,312 variable sites, i.e.,  $2.67 \times 10^6$  pairwise interactions). All 12 sequences with both PB1-317 and PB2-355 mutations also contained both NS1-92 and NS1-228 mutations. Hence, the model and wider results indicate that the particular polymerase and NS1 gene combination significantly affects HPAI H5N1 virulence in mammals and that a mutation pattern associated with virulence in mammals in the 1997 Hong Kong outbreak has reemerged among human H5N1 cases in Indonesia and China in 2006 and 2007.

The other mutations directly associated with virulence in the BGM are NS1-195T and HA-102V. These sites are not strongly linked in the model to the PB1, PB2, and NS1 sites discussed above and may represent a separate mechanism for enhanced virulence in mammals. NS1 sites S195, T197, and D92 are hydrogen bonded in the NS1 crystal structure (6) and make a cleft. It has been proposed that mutation D92E could alter the phosphorylation of this NS1 cleft and thus affect virulence (6, 44). The BGM results indicate that NS1-195 mutations also affect virulence, and it is postulated that this could be due to NS1-195 mutations changing the phosphorylation properties of NS1, although clearly more experimental work is

required to investigate this as a possible mechanism. Interestingly, the pattern of HA-102(86) V residues were completely correlated with the NS1-195 T in the 69 sequence data set. HA-102(86) is part of antigenic site E (83), so it is conceivable that a mutation here could reduce any cross-immunity that mammals may have against the virus. The correlation between NS1-195T and HA-102V is highly significant in the 579 full-genome sequence set ( $P \ll 10^{-6}$  for the Fisher Exact test with the Bonferroni correction), and there are 83 sequences with both mutations, recorded from 2004 onward, mostly in Thailand, Viet Nam, and Russia. HA-102 is a highly polymorphic site, with six different residues present in the 69-sequence set. Of these, V (present in 10 isolates) and A (present in 55 isolates) are associated with virulence and nonvirulence in mammals, respectively. The remaining four residues (I, P, S, and T) are present in one isolate each, making a virulence assignment on the basis of the 69 sequence data set alone difficult. In the 579-sequence data set, residues I, P, and S were rare and only present in 13 avian isolates in total. HA-102T, on the other hand, was present in 48 isolates: 44 of the 48 human isolates from Indonesia; a cat isolate from Indonesia, and 3 avian isolates from China. In the BGM, HA-102V was directly associated with virulence with a link probability of 0.6 to 0.7 (standard score = 3 to 4), and updating the binary coding scheme for HA-102 to include T as a "virulent" residue increases the link probability in the BGM to 0.8 to 0.9 (the standard score remains 3 to 4). Consequently, HA-102 V or T should be considered as a virulence marker.

At first sight, it is surprising that PB2-627 alone was not found to be directly associated to virulence in mammals because this mutation is thought to be involved in adaptation to mammalian hosts (73), has been previously reported as contributing to virulence in the 1997 Hong Kong outbreak (10, 24, 25), and was shown to be important for efficient respiratory tract growth in mammals (25). The sequence set used here spanned more than the well-studied 1997 Hong Kong outbreak and contained 6 PB2-627K virulent sequences (of 22 virulent sequences), and 1 PB2-627K nonvirulent sequence (of 47). The six virulent PB2-627K sequences were sampled in Hong Kong in 1997 (2), Thailand and Viet Nam in 2004 (3), and Germany in 2006 (1), whereas the single PB2-627K sequence classified as nonvirulent was from A/Turkey/65596/2006 (a human case in Turkey), which caused only mild symptoms in ferrets at a low dose (85). In one of the cross-validation data sets, the PB2-627K nonvirulent sequence was (randomly) excluded, and the resulting BGM had a PB2-627-virulence link probability of 0.6. However, using only sequences that were either very highly pathogenic or nonpathogenic (75% of cases, 52 sequences), no significant link was found between PB2-627K and virulence in mammals. On the other hand, since it has been suggested that a PB2-701D-to-N mutation may arise to compensate for the lack of the "virulent" PB2-627K residue in mammals (17, 38, 70), we combined the PB2-627 and PB2-701 sites, repeated the analysis, and found a significant direct association between the combined sites and virulence in mammals using BGMs (link probability of  $\geq 0.8$ , standard score = 3 to 4). Although the relationship between PB2-627K alone and virulence in mammals was not found to be significant in the 69 sequence data set, these results suggest that PB2-627K and PB2-701N might be directly associated with virulence, but more isolates with



these mutations and experimentally measured virulence in mammals would be required to detect the associations unambiguously. In addition, more isolates would allow the putative negative association between PB2-627K and PB2-701N to be evaluated.

Even though the uncorrected  $P$  value from the Fisher exact test for the association of NA stalk deletion and virulence in mammals was  $\sim 0.0003$ , the BGM results did not show a direct association. However, strong direct associations of NA stalk deletion with HA-172(156, glycosylation site) and HA-279(263, antigenic site) were inferred. HA and NA have complementary functions in viral replication (see Wagner et al. [78] for a review), and various studies have indicated that the effects of mutations in one can be compensated by mutations in the other (52, 54, 79); consequently, some HA-NA associations are to be expected. In particular, the NA stalk deletion-HA-172 glycosylation intersegment association (3) was confirmed by the BGM in the present study. Both the association of HA-172 and NA stalk deletion and the HA-279 and NA stalk deletion were significant in the 579-genome sequence set (Bonferroni-corrected Fisher exact test result of  $P < 10^{-6}$  for each). Neither the BGM nor the large sequence set provided any evidence that HA-172 and HA-279 were associated with each other (the corrected  $P$  value for the Fisher exact test on the large sequence set was 1), so we deduce that the NA stalk deletion is independently associated with both HA mutations, indicating a functional interaction between HA and NA.

An underlying issue with analysis of genetic determinants in influenza is the general lack of recombination within segments and consequent strong associations of amino acid variants over evolutionary time. Among the 228 sites with  $P \leq 0.05$  for association with virulence in the 69-sequence data set, the high degree of correlation in patterns of mutation is largely due to identity by descent of the viral genome segments. Thus, 34 sites have most or all of their "virulent" residues in the 11 sequences isolated in the 1997 Hong Kong outbreak, 15 sites on segment 8 have mostly "virulent" residues except in the Goose/Guangdong/96 and 1999-2001 NS allele B sequences (9, 21, 49), and 7 sites have most or all of their "virulent" residues in the 10 sequences from the 2004 Vietnam/Thailand outbreak.

To mitigate the potential bias in the results introduced by nonindependence, an alternative approach would be to code each sequence for the presence or absence of mutations compared to its immediate inferred ancestral sequence (using a phylogenetic tree and model of sequence evolution). The change in the viral phenotype from the ancestral sequences would also have to be inferred. However, although this approach was successfully used to find mutations in the HIV envelope protein (60, 61), it cannot be applied to a data set where each segment has its own evolutionary history (22, 77, 80) since reassortment of the internal genes with other subtypes means that no single tree is applicable to all eight segments. Recognizing the issue, we reduced potential biases incurred due to the relationships between the sequences by using a relatively wide range of isolates (spanning more than 10 years) and selection of sites with dissimilar mutation patterns across the strains.

We used several techniques to reduce other potential sources of error in the model. First, backed by simulation studies investigating the number of false-positive associations

as a function of data set size and number of sites, we limited the number of sites in our analysis to those with prior evidence from the literature and the strongest univariate associations to virulence in mammals. Second, we used cross-validation to reduce the effects of sequence misclassification (and sequence errors): the model presented here is an average over 10 subsamples of the data. Third, we used permutation tests to validate the significance of the associations against the null model of association by chance.

Using the model, we estimated the probability of virulence in mammals for strains containing different combinations of influencing sites. To evaluate the prediction performance of the model, we predicted each sequence in the cross-validation test data sets as virulent or nonvirulent on the basis of the combination of mutations it contained and compared the results to the original values. The high specificity (95%) indicates the model yields few false positives, but the lower sensitivity (70%) indicates that not all virulence determinants may have been captured as a result of the small size of the data set. Noise and experimental errors in the data are also expected to reduce the classification performance. The performance of the model compares well with other models relating viral mutations to phenotype, e.g., sensitivity in the range 40 to 100% (70% was typical) and a specificity of ca. 70% have been reported for HIV drug resistance models (42).

In conclusion, we have collated experimental data about the virulence of HPAI H5N1 strains in mammals and used it to identify its genetic basis by inferring a BGM of whole-genome mutations affecting the phenotype. We used data from strains isolated over a period of at least 10 years and analyzed the results obtained in 27 studies so our conclusions are general rather than specific to a single experimental system. The resulting statistical model reveals the polygenic nature of virulence in mammals: mutations at PB1-317, PB2-355, NS1-92, NS1-228, NS1-195 and HA-102(86) are directly associated with the trait. These sites split into two semi-independent groups, possibly indicating separate virulence mechanisms: PB1-317, PB2-355, NS1-92, and NS1-228 mutations were strongly correlated with each other, as were NS1-195 and HA-102(86). We also found strong correlation patterns in the PB1, PB2, and PA sites; the antigenic sites and a variable glycosylation site in HA; and the HA and NS sites. The results highlight the importance of the polymerase and nonstructural internal genes in addition to HA in determining the virulence of HPAI H5N1 in mammals and identify potential targets for intervention in mammalian H5N1 infections.

#### ACKNOWLEDGMENTS

This study was funded by the Biotechnology and Biological Sciences Research Council (Bb/B009670/1). In addition, this research was supported in part by the National Institutes of Health (AI43638, AI47745, and AI57167), the University of California Universitywide AIDS Research Program (grant number IS02-SD-701), and a University of California, San Diego, Center for AIDS Research/NIAID Developmental Award (AI36214).

#### REFERENCES

1. Alexander, D. J. 2000. A review of avian influenza in different bird species. *Vet. Microbiol.* 74:3-13.
2. Amonsin, A., T. Songserm, S. Chutinimitkul, R. Jam-on, N. Sae-Heng, N. Pariyothorn, S. Payungporn, A. Theamboonlers, and Y. Poovorawan. 2007. Genetic analysis of influenza A virus (H5N1) derived from domestic cat and dog in Thailand. *Arch. Virol.* 152:1925-1933.

3. Baigent, S. J., and J. W. McCauley. 2001. Glycosylation of haemagglutinin and stalk-length of neuraminidase combine to regulate the growth of avian influenza viruses in tissue culture. *Virus Res.* **79**:177–185.
4. Bao, Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman. 2008. The influenza virus resource at the National Center for Biotechnology Information. *J. Virol.* **82**:596–601.
5. Bender, C., H. Hall, J. Huang, A. Klimov, N. Cox, A. Hay, V. Gregory, K. Cameron, W. Lim, and K. Subbarao. 1999. Characterization of the surface proteins of influenza A (H5N1) viruses isolated from humans in 1997–1998. *Virology* **254**:115–123.
6. Bornholdt, Z. A., and B. V. V. Prasad. 2006. X-ray structure of influenza virus NS1 effector domain. *Nat. Struct. Mol. Biol.* **13**:559–560.
7. Bornholdt, Z. A., and B. V. V. Prasad. 2008. X-ray structure of NS1 from a highly pathogenic H5N1 influenza virus. *Nature* **456**:985–988.
8. Bright, R. A., T. M. Ross, K. Subbarao, H. L. Robinson, and J. M. Katz. 2003. Impact of glycosylation on the immunogenicity of a DNA-based influenza H5 HA vaccine. *Virology* **308**:270–278.
9. Cauthen, A. N., D. E. Swayne, S. Schultz-Cherry, M. L. Perdue, and D. L. Suarez. 2000. Continued circulation in china of highly pathogenic avian influenza viruses encoding the hemagglutinin gene associated with the 1997 H5N1 outbreak in poultry and humans. *J. Virol.* **74**:6592–6599.
10. Chen, H., R. A. Bright, K. Subbarao, C. Smith, N. J. Cox, J. M. Katz, and Y. Matsuoka. 2007. Polygenic virulence factors involved in pathogenesis of 1997 Hong Kong H5N1 influenza viruses in mice. *Virus Res.* **128**:159–163.
11. Chen, H., G. Deng, Z. Li, G. Tian, Y. Li, P. Jiao, L. Zhang, Z. Liu, R. G. Webster, and K. Yu. 2004. The evolution of H5N1 influenza viruses in ducks in southern China. *Proc. Natl. Acad. Sci. USA* **101**:10452–10457.
12. Chen, H., G. J. D. Smith, K. S. Li, J. Wang, X. H. Fan, J. M. Rayner, D. Vijaykrishna, J. X. Zhang, L. J. Zhang, C. T. Guo, C. L. Cheung, K. M. Xu, L. Duan, K. Huang, K. Qin, Y. H. C. Leung, W. L. Wu, H. R. Lu, Y. Chen, N. S. Xia, T. S. P. Naipospos, K. Y. Yuen, S. S. Hassan, S. Bahri, T. D. Nguyen, R. G. Webster, J. S. M. Peiris, and Y. Guan. 2006. Establishment of multiple sublineages of H5N1 influenza virus in Asia: implications for pandemic control. *Proc. Natl. Acad. Sci. USA* **103**:2845–2850.
13. Chen, H., G. J. D. Smith, S. Y. Zhang, K. Qin, J. Wang, K. S. Li, R. G. Webster, J. S. M. Peiris, and Y. Guan. 2005. Avian flu H5N1 virus outbreak in migratory waterfowl. *Nature* **436**:191–192.
14. Chen, R., and E. C. Holmes. 2006. Avian influenza virus exhibits rapid evolutionary dynamics. *Mol. Biol. Evol.* **23**:2336–2341.
15. Claas, E. C. J., A. D. M. E. Osterhaus, R. van Beek, J. C. De Jong, G. F. Rimmelzwaan, D. A. Senne, S. Krauss, K. F. Shortridge, and R. G. Webster. 1998. Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *Lancet* **351**:472–477.
16. Cooper, G. F., and E. Herskovits. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* **9**:309–347.
17. de Jong, M. D., C. P. Simmons, T. T. Thanh, V. M. Hien, G. J. D. Smith, T. N. B. Chau, D. M. Hoang, N. Van Vinh Chau, T. H. Khanh, V. C. Dong, P. T. Qui, B. Van Cam, D. Q. Ha, Y. Guan, J. S. M. Peiris, N. T. Chinh, T. T. Hien, and J. Farrar. 2006. Fatal outcome of human influenza A (H5N1) is associated with high viral load and hypercytokinemia. *Nat. Med.* **12**:1203–1207.
18. Friedman, N. 2004. Inferring cellular networks using probabilistic graphical models. *Science* **303**:799–805.
19. Friedman, N., and D. Koller. 2003. Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks. *Machine Learning* **50**:95–125.
20. Govorkova, E. A., J. E. Rehg, S. Krauss, H.-L. Yen, Y. Guan, M. Peiris, T. D. Nguyen, T. H. Hanh, P. Puthavathana, H. T. Long, C. Buranathai, W. Lim, R. G. Webster, and E. Hoffmann. 2005. Lethality to ferrets of H5N1 influenza viruses isolated from humans and poultry in 2004. *J. Virol.* **79**:2191–2198.
21. Guan, Y., J. S. M. Peiris, A. S. Lipatov, T. M. Ellis, K. C. Dyrting, S. Krauss, L. J. Zhang, R. G. Webster, and K. F. Shortridge. 2002. Emergence of multiple genotypes of H5N1 avian influenza viruses in Hong Kong SAR. *Proc. Natl. Acad. Sci. USA* **99**:8950–8955.
22. Guan, Y., L. L. M. Poon, C. Y. Cheung, T. M. Ellis, W. Lim, A. S. Lipatov, K. H. Chan, K. M. Sturm-Ramirez, C. L. Cheung, Y. H. C. Leung, K. Y. Yuen, R. G. Webster, and J. S. M. Peiris. 2004. H5N1 influenza: a protean pandemic threat. *Proc. Natl. Acad. Sci. USA* **101**:8156–8161.
23. Ha, Y., D. J. Stevens, J. J. Skehel, and D. C. Wiley. 2001. X-ray structures of H5 avian and H9 swine influenza virus hemagglutinins bound to avian and human receptor analogs. *Proc. Natl. Acad. Sci. USA* **98**:11181–11186.
24. Hatta, M., P. Gao, P. Halfmann, and Y. Kawaoka. 2001. Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. *Science* **293**:1840–1842.
25. Hatta, M., Y. Hatta, J. H. Kim, S. Watanabe, K. Shinya, T. Nguyen, P. S. Lien, Q. M. Le, and Y. Kawaoka. 2007. Growth of H5N1 influenza A viruses in the upper respiratory tracts of mice. *PLoS Pathog.* **3**:e133.
26. Heikkinen, L. S., A. Kazlauskas, K. Melen, R. Wagner, T. Ziegler, I. Julkunen, and K. Saksela. 2008. Avian and 1918 Spanish influenza A virus NS1 proteins bind to Crk/CrkL SH3 domains to activate host cell signalling. *J. Biol. Chem.* **283**:5719–5727.
27. Horimoto, T., and Y. Kawaoka. 1994. Reverse genetics provides direct evidence for a correlation of hemagglutinin cleavability and virulence of an avian influenza A virus. *J. Virol.* **68**:3120–3128.
28. Hulse-Post, D. J., J. Franks, K. Boyd, R. Salomon, E. Hoffmann, H. L. Yen, R. J. Webby, D. Walker, T. D. Nguyen, and R. G. Webster. 2007. Molecular changes in the polymerase genes (PA and PB1) associated with high pathogenicity of H5N1 influenza virus in mallard ducks. *J. Virol.* **81**:8515–8524.
29. Hulse-Post, D. J., K. M. Sturm-Ramirez, J. Humberd, P. Seiler, E. A. Govorkova, S. Krauss, C. Scholtissek, P. Puthavathana, C. Buranathai, T. D. Nguyen, H. T. Long, T. S. P. Naipospos, H. Chen, T. M. Ellis, Y. Guan, J. S. M. Peiris, and R. G. Webster. 2005. Role of domestic ducks in the propagation and biological evolution of highly pathogenic H5N1 influenza viruses in Asia. *Proc. Natl. Acad. Sci. USA* **102**:10682–10687.
30. Ito, T., Y. Suzuki, A. Takada, A. Kawamoto, K. Otsuki, H. Masuda, M. Yamada, T. Suzuki, H. Kida, and Y. Kawaoka. 1997. Differences in sialic acid-galactose linkages in the chicken egg amnion and allantois influence human influenza virus receptor specificity and variant selection. *J. Virol.* **71**:3357–3362.
31. Jackson, D., M. J. Hossain, D. Hickman, D. R. Perez, and R. A. Lamb. 2008. A new influenza virus virulence determinant: the NS1 protein four C-terminal residues modulate pathogenicity. *Proc. Natl. Acad. Sci. USA* **105**:4381–4386.
32. Jansen, R., N. Lan, J. Qian, and M. Gerstein. 2002. Integration of genomic datasets to predict protein complexes in yeast. *J. Struct. Funct. Genomics* **2**:71–81.
33. Jiao, P., G. Tian, Y. Li, G. Deng, Y. Jiang, C. Liu, W. Liu, Z. Bu, Y. Kawaoka, and H. Chen. 2008. A single-amino-acid substitution in the NS1 protein changes the pathogenicity of H5N1 avian influenza viruses in mice. *J. Virol.* **82**:1146–1154.
34. Katz, J. M., X. Lu, T. M. Tumpey, C. B. Smith, M. W. Shaw, and K. Subbarao. 2000. Molecular correlates of influenza A H5N1 virus pathogenesis in mice. *J. Virol.* **74**:10807–10810.
35. Keawcharoen, J., K. Oraveerakul, T. Kuiken, R. A. Fouchier, A. Amonsin, S. Payungporn, S. Noppornpanth, S. Wattanadorn, A. Theambooniers, R. Tantilertcharoen, R. Pattanarangsarn, N. Arya, P. Ratanakorn, D. M. Osterhaus, and Y. Poovorawan. 2004. Avian influenza H5N1 in tigers and leopards. *Emerg. Infect. Dis.* **10**:2189–2191.
36. Kuiken, T., G. Rimmelzwaan, D. van Riel, G. van Amerongen, M. Baars, R. Fouchier, and A. Osterhaus. 2004. Avian H5N1 influenza in cats. *Science* **306**:241.
37. Lauritzen, S. L., and N. A. Sheehan. 2003. Graphical models for genetic analyses. *Stat. Sci.* **18**:489–514.
38. Le, Q. M., Y. Sakai-Tagawa, M. Ozawa, M. Ito, and Y. Kawaoka. 2009. Selection of H5N1 influenza virus PB2 during replication in humans. *J. Virol.* **83**:5278–5281.
39. Lee, C.-W., D. L. Suarez, T. M. Tumpey, H.-W. Sung, Y.-K. Kwon, Y.-J. Lee, J.-G. Choi, S.-J. Joh, M.-C. Kim, E.-K. Lee, J.-M. Park, X. Lu, J. M. Katz, E. Spackman, D. E. Swayne, and J.-H. Kim. 2005. Characterization of highly pathogenic H5N1 avian influenza A viruses isolated from South Korea. *J. Virol.* **79**:3692–3702.
40. Lee, I., S. V. Date, A. T. Adai, and E. M. Marcotte. 2004. A probabilistic functional network of yeast genes. *Science* **306**:1555–1558.
41. Lee, M. S., M. C. Deng, Y. J. Lin, C. Y. Chang, H. K. Shieh, J. Z. Shiau, and C. C. Huang. 2007. Characterization of an H5N1 avian influenza virus from Taiwan. *Vet. Microbiol.* **124**:193–201.
42. Leigh Brown, A. J., S. D. W. Frost, B. Good, E. S. Daar, V. Simon, M. Markowitz, A. C. Collier, E. Connick, B. Conway, J. B. Margolick, J.-P. Routy, J. Corbeil, N. S. Hellmann, D. D. Richman, and S. J. Little. 2004. Genetic basis of hypersusceptibility to protease inhibitors and low replicative capacity of human immunodeficiency virus type 1 strains in primary infection. *J. Virol.* **78**:2242–2246.
43. Li, K. S., Y. Guan, J. Wang, G. J. D. Smith, K. M. Xu, L. Duan, A. P. Rahardjo, P. Puthavathana, C. Buranathai, T. D. Nguyen, A. T. S. Estepangestie, A. Chaisingh, P. Auewarakul, H. T. Long, N. T. H. Hanh, R. J. Webby, L. L. M. Poon, H. Chen, K. F. Shortridge, K. Y. Yuen, R. G. Webster, and J. S. M. Peiris. 2004. Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature* **430**:209–213.
44. Li, M., and B. Wang. 2007. Homology modeling and examination of the effect of the D92E mutation on the H5N1 nonstructural protein NS1 effector domain. *J. Mol. Model.* **13**:1237–1244.
45. Li, Z., Y. Jiang, P. Jiao, A. Wang, F. Zhao, G. Tian, X. Wang, K. Yu, Z. Bu, and H. Chen. 2006. The NS1 gene contributes to the virulence of H5N1 avian influenza viruses. *J. Virol.* **80**:11115–11123.
46. Lipatov, A. S., S. Andreansky, R. J. Webby, D. J. Hulse, J. E. Rehg, S. Krauss, D. R. Perez, P. C. Doherty, R. G. Webster, and M. Y. Sangster. 2005. Pathogenesis of Hong Kong H5N1 influenza virus NS gene reassortants in mice: the role of cytokines and B- and T-cell responses. *J. Gen. Virol.* **86**:1121–1130.
47. Lipatov, A. S., S. Krauss, Y. Guan, M. Peiris, J. E. Rehg, D. R. Perez, and R. G. Webster. 2003. Neurovirulence in mice of H5N1 influenza virus genotypes isolated from Hong Kong poultry in 2001. *J. Virol.* **77**:3816–3823.
48. Long, J.-X., D.-X. Peng, Y.-L. Liu, Y.-T. Wu, and X.-F. Liu. 2008. Virulence

- of H5N1 avian influenza virus enhanced by a 15-nucleotide deletion in the viral nonstructural gene. *Virus Genes* **36**:471–478.
49. Ludwig, S., U. Schultz, J. Mandler, W. M. Fitch, and C. Scholtissek. 1991. Phylogenetic relationship of the nonstructural (NS) genes of influenza A viruses. *Virology* **183**:566–577.
  50. Maines, T. R., X. H. Lu, S. M. Erb, L. Edwards, J. Guarner, P. W. Greer, D. C. Nguyen, K. J. Szretter, L.-M. Chen, P. Thawatsupha, M. Chittaganpitch, S. Waicharoen, D. T. Nguyen, T. Nguyen, H. H. T. Nguyen, J.-H. Kim, L. T. Hoang, C. Kang, L. S. Phuong, W. Lim, S. Zaki, R. O. Donis, N. J. Cox, J. M. Katz, and T. M. Tumpey. 2005. Avian influenza (H5N1) viruses isolated from humans in Asia in 2004 exhibit increased virulence in mammals. *J. Virol.* **79**:11788–11800.
  51. Marcotte, E. M., I. Xenarios, and D. Eisenberg. 2001. Mining literature for protein-protein interactions. *Bioinformatics* **17**:359–363.
  52. Matrosovich, M., N. Zhou, Y. Kawaoka, and R. Webster. 1999. The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distinguishable properties. *J. Virol.* **73**:1146–1155.
  53. Min, J.-Y., S. Li, G. C. Sen, and R. M. Krug. 2007. A site on the influenza A virus NS1 protein mediates both inhibition of PKR activation and temporal regulation of viral RNA synthesis. *Virology* **363**:236–243.
  54. Mitnaul, L. J., M. N. Matrosovich, M. R. Castrucci, A. B. Tuzikov, N. V. Bovin, D. Kobasa, and Y. Kawaoka. 2000. Balanced hemagglutinin and neuraminidase activities are critical for efficient replication of influenza A virus. *J. Virol.* **74**:6015–6020.
  55. Mushtaq, M., H. Juan, P. Jiang, Y. Li, T. Li, Y. Du, and M. Mukhtar. 2008. Complete genome analysis of a highly pathogenic H5N1 influenza A virus isolated from a tiger in China. *Arch. Virol.* **153**:1569–1574.
  56. Myers, C., D. Barrett, M. Hibbs, C. Huttenhower, and O. Troyanskaya. 2006. Finding function: evaluation methods for functional genomic data. *BMC Genomics* **7**:187.
  57. Obenauer, J. C., J. Denson, P. K. Mehta, X. Su, S. Mukatira, D. B. Finkelstein, X. Xu, J. Wang, J. Ma, Y. Fan, K. M. Rakestraw, R. G. Webster, E. Hoffmann, S. Krauss, J. Zheng, Z. Zhang, and C. W. Naeve. 2006. Large-scale sequence analysis of avian influenza isolates. *Science* **311**:1576–1580.
  58. Olsen, B., V. J. Munster, A. Wallensten, J. Waldenstrom, A. D. M. E. Osterhaus, and R. A. M. Fouchier. 2006. Global patterns of influenza A virus in wild birds. *Science* **312**:384–388.
  59. Pearl, J. 1986. Fusion, propagation, and structuring in belief networks. *Artif. Intell.* **29**:241–288.
  60. Poon, A. F., F. I. Lewis, S. D. Frost, and S. L. Kosakovsky Pond. 2008. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. *Bioinformatics* **24**:1949–1950.
  61. Poon, A. F., F. I. Lewis, S. L. Pond, and S. D. Frost. 2007. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. *PLoS Comput. Biol.* **3**:e231.
  62. Rott, R. 1980. Genetic determinants for infectivity and pathogenicity of influenza viruses. *Phil. Trans. R. Soc. London B Biol. Sci.* **288**:393–399.
  63. Rott, R. 1992. The pathogenic determinant of influenza virus. *Vet. Microbiol.* **33**:303–310.
  64. Salomon, R., J. Franks, E. A. Govorkova, N. A. Ilyushina, H.-L. Yen, D. J. Hulse-Post, J. Hummerd, M. Trichet, J. E. Rehg, R. J. Webby, R. G. Webster, and E. Hoffmann. 2006. The polymerase complex genes contribute to the high virulence of the human H5N1 influenza virus isolate A/Vietnam/1203/04. *J. Exp. Med.* **203**:689–697.
  65. Searls, D. B. 2005. Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* **4**:45–58.
  66. Sebastiani, P., M. F. Ramoni, V. Nolan, C. T. Baldwin, and M. H. Steinberg. 2005. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nat. Genet.* **37**:435–440.
  67. Seo, S. H., E. Hoffmann, and R. G. Webster. 2002. Lethal H5N1 influenza viruses escape host antiviral cytokine responses. *Nat. Med.* **8**:950–954.
  68. Smith, G. J. D., T. S. P. Naipospos, T. D. Nguyen, M. D. de Jong, D. Vijaykrishna, T. B. Usman, S. S. Hassan, T. V. Nguyen, T. V. Dao, N. A. Bui, Y. H. C. Leung, C. L. Cheung, J. M. Rayner, J. X. Zhang, L. J. Zhang, L. L. M. Poon, K. S. Li, V. C. Nguyen, T. T. Hien, J. Farrar, R. G. Webster, H. Chen, J. S. M. Peiris, and Y. Guan. 2006. Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam. *Virology* **350**:258–268.
  69. Stark, C., B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. 2006. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**:D535–D539.
  70. Steel, J., A. C. Lowen, S. Mubareka, and P. Palese. 2009. Transmission of influenza virus in a mammalian host is increased by PB2 amino acids 627K or 627E/701N. *PLoS Pathog.* **5**:e1000252.
  71. Sturm-Ramirez, K. M., T. Ellis, B. Bousfield, L. Bissett, K. Dyrting, J. E. Rehg, L. Poon, Y. Guan, M. Peiris, and R. G. Webster. 2004. Reemerging H5N1 influenza viruses in Hong Kong in 2002 are highly pathogenic to ducks. *J. Virol.* **78**:4892–4901.
  72. Suarez, D. L., M. L. Perdue, N. Cox, T. Rowe, C. Bender, J. Huang, and D. E. Swayne. 1998. Comparisons of highly virulent H5N1 influenza A viruses isolated from humans and chickens from Hong Kong. *J. Virol.* **72**:6678–6688.
  73. Subbarao, E. K., W. London, and B. R. Murphy. 1993. A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. *J. Virol.* **67**:1761–1764.
  74. Subbarao, K., A. Klimov, J. Katz, H. Regnery, W. Lim, H. Hall, M. Perdue, D. Swayne, C. Bender, J. Huang, M. Hemphill, T. Rowe, M. Shaw, X. Xu, K. Fukuda, and N. Cox. 1998. Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness. *Science* **279**:393–396.
  75. Subbarao, K., and M. W. Shaw. 2000. Molecular aspects of avian influenza (H5N1) viruses isolated from humans. *Rev. Med. Virol.* **10**:337–348.
  76. Thiry, E., A. Zicola, D. Addie, H. Egberink, K. Hartmann, H. Lutz, H. Poulet, and M. C. Horzinek. 2007. Highly pathogenic avian influenza H5N1 virus in cats and other carnivores. *Vet. Microbiol.* **122**:25–31.
  77. Vijaykrishna, D., J. Bahl, S. Riley, L. Duan, J. X. Zhang, H. Chen, J. S. M. Peiris, G. J. D. Smith, and Y. Guan. 2008. Evolutionary dynamics and emergence of panzootic H5N1 influenza viruses. *PLoS Pathog.* **4**:e1000161.
  78. Wagner, R., M. Matrosovich, and H. D. Klenk. 2002. Functional balance between haemagglutinin and neuraminidase in influenza virus infections. *Rev. Med. Virol.* **12**:159–166.
  79. Wagner, R., T. Wolff, A. Herwig, S. Pleschka, and H. D. Klenk. 2000. Interdependence of hemagglutinin glycosylation and neuraminidase as regulators of influenza virus growth: a study by reverse genetics. *J. Virol.* **74**:6316–6323.
  80. Wan, X.-F., T. Nguyen, C. T. Davis, C. B. Smith, Z.-M. Zhao, M. Carrel, K. Inui, H. T. Do, D. T. Mai, S. Jadhao, A. Balish, B. Shu, F. Luo, M. Emch, Y. Matsuoka, S. E. Lindstrom, N. J. Cox, C. V. Nguyen, A. Klimov, and R. O. Donis. 2008. Evolution of highly pathogenic H5N1 avian influenza viruses in Vietnam between 2001 and 2007. *PLoS ONE* **3**:e3462.
  81. Webster, R. G., S. Krauss, D. Hulse-Post, and K. Sturm-Ramirez. 2007. Evolution of influenza A viruses in wild birds. *J. Wildl. Dis.* **43**:S1–S6.
  82. WHO/OIE/FAO H5N1 Evolution Working Group. 2008. Toward a unified nomenclature system for highly pathogenic avian influenza virus (H5N1). *Emerg. Infect. Dis.* **14**:e1.
  83. Wu, W. L., Y. Chen, P. Wang, W. Song, S.-Y. Lau, J. M. Rayner, G. J. D. Smith, R. G. Webster, J. S. M. Peiris, T. Lin, N. Xia, Y. Guan, and H. Chen. 2008. Antigenic profile of avian H5N1 viruses in Asia from 2002 to 2007. *J. Virol.* **82**:1798–1807.
  84. Xu, X., K. Subbarao, N. J. Cox, and Y. Guo. 1999. Genetic characterization of the pathogenic influenza A/Goose/Guangdong/1/96 (H5N1) virus: similarity of its hemagglutinin gene to those of H5N1 viruses from the 1997 outbreaks in Hong Kong. *Virology* **261**:15–19.
  85. Yen, H. L., A. S. Lipatov, N. A. Ilyushina, E. A. Govorkova, J. Franks, N. Yilmaz, A. Douglas, A. Hay, S. Krauss, J. E. Rehg, E. Hoffmann, and R. G. Webster. 2007. Inefficient transmission of H5N1 influenza viruses in a ferret contact model. *J. Virol.* **81**:6890–6898.
  86. Zitzow, L. A., T. Rowe, T. Morken, W. J. Shieh, S. Zaki, and J. M. Katz. 2002. Pathogenesis of avian influenza A (H5N1) viruses in ferrets. *J. Virol.* **76**:4420–4429.